Exponential Weighting: Effects on Precision

Ludger Hentschel

December 9, 2024

Abstract
I explore the cost of exponentially weighted estimates in terms of reduced precision.

When the data generating process changes gradually over time, I often use exponentially weighted estimators in the hope of obtaining less biased estimates. Since the weights reduce the effective sample size, the estimators are less precise than equally weighted estimates.

I show that the lost precision, even for independent and identically distributed samples, is moderate when the half-life of the weights is similar to the sample length.

If the half-life of the weights is much smaller than the sample length, exponential weighting potentially sacrifices a great deal of precision. In those cases, especially, we should ensure that the weighting is appropriate for the data generating process and not just an *ad hoc* weighting to guard against unspecified changes.

www.ludgerhentschel.com ludger@ludgerhentschel.com

Contents

1 Introduction

We frequently apply exponential weights during estimation to give more weight to recent data and less weight to older data. Often, the main intent is to allow for unspecified, gradual changes in the data over time.

By design, exponential weighting is effective when data changes over time. We can show that exponential weighting (or smoothing) produces optimal forecasts when the data follow an ARIMA(0, 1, 1) process.[1] This time-series behavior, or something close to it, is common in many financial data series like return volatilities or asset prices.[2]

Exponential weighting is often applied in many other situations, where the data are not especially close to ARIMA(0, 1, 1). The objective is to accommodate unspecified, gradual changes over time without a specific claim that exponential weighting is statistically optimal. This may appear satisfactory if we have a vague sense of a benefit and don't track the costs. For stationary data with much less persistence than ARIMA(0, 1, 1), weighting comes at a cost of lower precision for the estimates. In these situations, there is a trade-off between allowing for gradual change (potentially less bias) and achieving higher precision (lower variance) that we should consider.

The objective of this note is to quantify the loss in precision due to exponential weighting. The cost of weighting in terms of lower precision compared to unweighted estimates increases as we reduce the half-life and as we increase the length of the sample. Both effects result in lower weights for older data.

Figure 1 illustrates the variance ratio of exponentially weighted estimates compared to equally weighted estimates. This is the worst case loss of efficiency, which occurs when exponential weighting is applied unnecessarily to independent and identically distributed (i.i.d.) data.

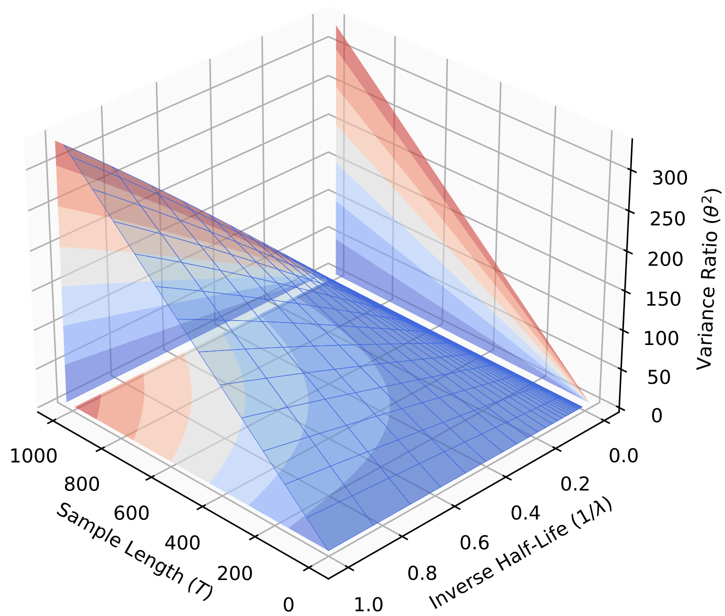
The benefit of *ad hoc* weighting is challenging to quantify in general, other than recognizing that exponentially weighting can be effective in the presence of gradual changes. The variance ratio can serve as a guide when we are trying to balance a rough sense of the benefits from exponential weighting against an approximate loss in efficiency. At a minimum, the variance ratio shows when the loss of efficiency is so large that we should have a clear justification for incurring this cost.

I derive the ratio of variances for exponentially weighted and equally weighted estimates of the mean. If the sample spans T periods and we apply

[1] See Muth (1960), for example.

[2] For example, see Bollerslev (1986) or Hentschel (1995) for time-series properties of return volatility. For example, see Fama (1965) or Lo and MacKinlay (1988) for time-series properties of asset prices.

Figure 1: Relative Precision



The figure shows the variance ratio θ^2 for exponentially weighted and equally weighted estimates. The relative precision depends on the sample length T and the half-life of the exponential weights λ. For clarity, the graph shows the inverse half-life, $1/\lambda$. The sample length and half-life must be measured in the same units but it does not matter whether those units are seconds, days, months, or years.

exponential weights with a half-life of λ, the variance of the exponentially weighted estimate is approximately $0.35\ T/\lambda$ times the variance of an equally weighted estimate. (For $T/\lambda > 3$.)

Although I focus on precision in estimates of a mean, the same analysis applies to a broad range of estimates with "root-N" standard errors. We can also apply similar analysis to other weighting schemes. However, the loss of efficiency depends on the pattern of the weights.

The remainder of the note provides the analysis supporting these conclusions.

2 Weights and Precision

Take a sample with n observations per period and T periods. This corresponds to $N = nT$ total observations. Let each observation come from the same distribution with a mean μ and variance σ^2. For i.i.d. data like this, weighting does not provide any benefit but we can compute the costs.

The interest in understanding the costs comes from the common practice of applying exponential weights even when they are not statistically optimal. We often give more weight to recent observations in the belief that they are more representative of the near future without specifying the time-series process that would lead to such changes. There is vague sense of a benefit from weighting the observations. The main objective here is to balance this with approximate costs of weighting the estimate. It is natural to estimate the maximum costs for the case where the weighting is unnecessary: i.i.d data.

2.1 Equally Weighted Estimates

The variance for the simple mean is

$$\begin{aligned} \sigma_0^2 &= \sum_{t=0}^{T-1} \sum_{i=1}^{N} \frac{1}{(nT)^2} \sigma^2 \\ &= \frac{1}{nT}\, \sigma^2. \end{aligned} \tag{1}$$

Importantly, this estimator has minimum variance among all linear unbiased estimators of the mean for i.i.d. data.[3] Since the exponentially weighted mean is also a linear estimator, this says that the exponentially weighted estimator has higher variance for i.i.d. data. The main question I answer is "How much higher?"

2.2 Exponentially Weighted Estimates

The variance for the exponentially weighted mean depends on the exponential weights. For an exponential coefficient $0 \leq \phi \leq 1$, the weights for all periods with lag t, $0 \leq t \leq T-1$, are

$$w_t = \frac{1-\phi}{1-\phi^T} \left\{ 1, \phi, \phi^2, \phi^3, \ldots, \phi^t, \ldots, \phi^{T-1} \right\}. \tag{2}$$

These weights are scaled so the sum of T weights equals 1.

For i.i.d. data, the variance of the exponentially weighted mean is

$$\begin{aligned} \sigma_1^2 &= \sum_{t=0}^{T-1} w_t^2 \sum_{i=1}^{n} \frac{1}{n^2}\, \sigma^2 \\ &= \frac{1}{n} \frac{(1-\phi)^2}{(1-\phi^T)^2} \sum_{t=0}^{T-1} \phi^{2t} \sigma^2 \end{aligned}$$

[3] This follows from the Gauss-Markov theorem. See the discussion in Chow (1983), for example. If the data are normally distributed, the sample mean has minimum variance among all unbiased estimators, linear or not. See the discussion in Hayashi (2000), for example.

$$
\begin{aligned}
&= \frac{1}{n} \frac{(1-\phi)^2}{(1-\phi^T)^2} \frac{(1-\phi^{2T})}{(1-\phi^2)} \sigma^2 \\
&= \frac{1}{n} \frac{1-\phi}{1+\phi} \frac{1+\phi^T}{1-\phi^T} \sigma^2 \\
&= \frac{1}{n\tau(\phi, T)} \sigma^2, \qquad (3)
\end{aligned}
$$

where

$$\tau(\phi, T) = \frac{1+\phi}{1-\phi} \frac{1-\phi^T}{1+\phi^T}. \qquad (4)$$

The factor τ measures the effective length of the exponentially weighted sample and $1/\tau$ plays the same role as $1/T$ for the equally weighted variance. We can confirm that τ converges to T as ϕ approaches 1.[4]

2.3 Relative Precision

The ratio of the variances is equal to the full sample length T divided by the effective sample length $\tau(\phi, T)$

$$
\begin{aligned}
\theta^2(\phi, T) &= \frac{\sigma_1^2}{\sigma_0^2} \\
&= T \frac{1-\phi}{1+\phi} \frac{1+\phi^T}{1-\phi^T} \\
&= \frac{T}{\tau(\phi, T)}. \qquad (5)
\end{aligned}
$$

We often describe exponential weights with reference to their half-life. The exponential coefficient ϕ is related to the half-life λ of the exponential weights via

$$\phi = \left(\frac{1}{2}\right)^{1/\lambda}. \qquad (6)$$

[4]We can rearrange terms, so that

$$
\begin{aligned}
\lim_{\phi \to 1} \tau(\phi, T) &= \lim_{\phi \to 1} \frac{1+\phi}{1+\phi^T} \lim_{\phi \to 1} \frac{1-\phi^T}{1-\phi} \\
&= \lim_{\phi \to 1} \frac{1-\phi^T}{1-\phi}
\end{aligned}
$$

since the first term converges to 1. The second ratio converges to 0 in both the numerator and denominator. Using L'Hôpital's rule, we can find this limit as

$$
\begin{aligned}
\lim_{\phi \to 1} \frac{1-\phi^T}{1-\phi} &= \lim_{\phi \to 1} \frac{-T\phi^{T-1}}{-1} \\
&= T.
\end{aligned}
$$

It is not illuminating to restate the variance ratio θ^2 or the effective sample length τ in terms of the half-life λ instead of the exponential coefficient ϕ. But it is easier to derive additional results in terms of the half-life.

Figure 1 shows the variance ratio for different combinations of sample lengths T and half-lives λ. To make the relation ship closer to linear and the graph clearer, the figure uses the inverse of the half-life, $1/\lambda$. The three-dimensional surface shows how the variance ratio θ^2 depends on the sample length T and the inverse half-life $1/\lambda$. In addition, the figure projects the surface onto the x, y, and z planes.

The figure shows that the variance ratio θ^2 is very large when the sample length is large and the half-life for the exponential weights is much smaller. Unless there are clear benefits in applying exponential weighting, we should carefully consider these costs in the precision of the estimates.

From equation (5), we can see that the variance ratio becomes infinite as the sample increases in length T and we hold ϕ constant. The variance of the equally weighted estimate converges to zero, while the variance of the exponentially weighted estimate converges to $(1/n)\ (1-\phi)/(1+\phi)\ \sigma^2 \approx (1/2n)\ (1-\phi)\ \sigma^2$. Similarly, we can see that the variance ratio converges to T as ϕ goes to zero and we hold the sample length T constant. In this case, the variance of the exponentially weighted average converges to σ^2/n, as all of the weight concentrates on the first period.[5]

The projections in figure 1 show that the variance ratio is approximately linear in T when we hold $1/\lambda$ constant, and vice versa. The z projection suggests that the variance ratio varies with the product of T and $1/\lambda$.

Sample Length

I have shown that relative precision is equal to the ratio of the actual sample length and the effective sample length and that the effective sample length is an increasing function of the exponential half-life.

An important maintained assumption is that T is the length of the available sample. To accommodate unspecified, gradual changes, we sometimes use only more recent observations, truncating the sample at some lag. This is another way of weighting observations: Equal weight to recent observations, zero weight to distant observations. The results presented here are meant to apply to the full sample length T, not a preselected subsample.

The loss of precision for truncated samples is more obvious than that for exponentially weighted samples. If we select a fraction α, $0 \leq \alpha \leq 1$, of the available sample, the effective sample length is αT. The variance ratio

[5] In continuous time, the variance ratio increases without limit as ϕ goes to zero since the weight concentrates on ever smaller fractions of the sample. See the appendix for details.

is equal to the full sample length divided by the effective sample length, $\theta^2 = T/(\alpha T) = 1/\alpha \geq 1$.

Estimates over rolling windows truncate the sample. Each estimate is based on a truncated sample. The sample gradually changes over time and we eventually use all observations. But no rolling window estimate uses the whole sample.

2.4 Approximations

Following the suggestion in figure 1, we can show that relative precision is approximately proportional to the product of T and $1/\lambda$.

We can approximate τ by assuming that ϕ is close to 1 but T is large enough to make ϕ^T essentially 0. Linearizing the ratio $(1-\phi)/(1+\phi)$ via a first-order Taylor series around $\phi = 1$, we find that

$$\theta^2(\phi, T) \approx \frac{T}{2}\,(1-\phi). \tag{7}$$

Substituting ϕ based on equation (6) yields

$$\theta^2(\lambda, T) \approx \frac{T}{2}\left(1 - \left(\frac{1}{2}\right)^{1/\lambda}\right). \tag{8}$$

Finally, taking a Taylor series expansion of $(1/2)^{1/\lambda}$ around $1/\lambda = 0$, which is consistent with $\phi = 1$, we find

$$\begin{aligned}
\theta^2(\lambda, T) &\approx \frac{T}{2}\left(1 - \left(1 + \frac{\ln(1/2)}{\lambda}\right)\right) \\
&\approx \frac{T}{\lambda}\,\frac{\ln(2)}{2} \qquad (9) \\
&\approx 0.35\,\frac{T}{\lambda}. \qquad (10)
\end{aligned}$$

In this approximate sense, the variance ratio is proportional to the ratio of sample length T and half-life λ. If the ratio T/λ is not "too large", then the relative inefficiency of the exponentially weighted estimates may be acceptable, even if exponential weighting is not optimal for the data generating process.

The approximation is a crude guide since we can immediately see that it breaks down for small ratios T/λ. The variance ratio must always be greater than 1, even for $T < \lambda$, contrary to what the approximation might suggest. Unfortunately, no proportional approximation can obey a lower bound like this. While higher-order approximations may provide numerically superior answers over a wider range of values, they sacrifice intuitive simplicity.[6]

Table 1: Variance Ratios

T (Years)	T	λ	T/λ	θ^2	$0.35\ T/\lambda$	θ
Panel A: 50 Years of Monthly Data						
50	600	2.40	250	86.05	87.50	9.28
50	600	4.80	125	43.25	43.75	6.58
50	600	12.00	50	17.32	17.50	4.16
50	600	30.00	20	6.93	7.00	2.63
50	600	60.00	10	3.47	3.50	1.86
50	600	120.00	5	1.84	1.75	1.36
50	600	150.00	4	1.57	1.40	1.25
50	600	200.00	3	1.34	1.05	1.16
50	600	300.00	2	1.16	0.70	1.07
50	600	600.00	1	1.04	0.35	1.02
Panel B: 10 Years of Daily Data						
10	2,500	10.00	250	86.61	87.50	9.31
10	2,500	20.00	125	43.32	43.75	6.58
10	2,500	50.00	50	17.33	17.50	4.16
10	2,500	125.00	20	6.93	7.00	2.63
10	2,500	250.00	10	3.47	3.50	1.86
10	2,500	500.00	5	1.84	1.75	1.36
10	2,500	625.00	4	1.57	1.40	1.25
10	2,500	833.33	3	1.34	1.05	1.16
10	2,500	1,250.00	2	1.16	0.70	1.07
10	2,500	2,500.00	1	1.04	0.35	1.02
Panel C: 1 Year of Daily Data						
1	250	1.00	250	83.33	87.50	9.13
1	250	2.00	125	42.89	43.75	6.55
1	250	5.00	50	17.30	17.50	4.16
1	250	12.50	20	6.93	7.00	2.63
1	250	25.00	10	3.47	3.50	1.86
1	250	50.00	5	1.84	1.75	1.36
1	250	62.50	4	1.57	1.40	1.25
1	250	83.33	3	1.34	1.05	1.16
1	250	125.00	2	1.16	0.70	1.07
1	250	250.00	1	1.04	0.35	1.02

The table compares the precision of exponentially weighted and equally weighted estimates for a range of sample sizes T and exponential half-lives λ.

The exponentially weighted estimate has variance θ^2 times the variance of the equally weighted estimate.

The table also shows the approximate variance ratio, $\theta^2 \approx 0.35\ T/\lambda$. The last column shows the ratio of standard errors, θ.

[6]In particular, a second-order approximation is not materially more accurate.

2.5 Examples

We can investigate the variance ratio and the linear approximation for plausible values of sample lengths and half-lives. Table 1 shows the variance ratio and its linear approximation for several combinations of T and λ. For practical purposes, it may be easiest to interpret both values in units of months or days. For financial data, there are about 250 trading days in a calendar year.

In addition to the variance ratio, the table also show its square root. We can interpret this as the ratio of standard errors for the estimates. This is an equivalent measure of relative precision but based on the same units as the estimate.

For ratios $T/\lambda < 5$, the reduced precision due to exponential weighting is moderate. For ratios much larger than 10, however, exponential weighting can induce a large loss in precision. In these cases, especially, we should be clear about the expected benefits in order to justify the loss in precision.

For ratios $T/\lambda > 4$, the approximation $\theta^2 \approx 0.35\ T/\lambda$, provides a useful guide to the true variance ratio.

Careful reading of table 1 reveals that the relative precision depends on T. This happens because θ^2 converges to T as the half-life λ goes to zero. For $\lambda < 1$, the approximation becomes less accurate. We rarely care about these cases and we know that the loss of efficiency, θ^2, is large.

The combination of these results suggests that we may be willing to accept the loss of precision from exponentially weighting when $T/\lambda < 4$, even if the benefits are vague, and to use the the approximation $\theta^2 \approx 0.35\ T/\lambda$ for the variance ratio to gauge the loss of precision otherwise.

3 Summary

We often apply exponential weights in estimation in order to allow for gradual, unspecified changes over time. Only in special cases do we think that such weighting is close to optimal, given the time-series process for the underlying data. In many cases, exponential weighting creates a loss of efficiency. I quantify the loss of precision in estimates due to exponentially weighting compared to equally weighting. This is the worst case loss, which occurs when exponential weighting is applied unnecessarily to i.i.d. data.

The results show that the loss in efficiency is fairly small when the sample length T is less than 5 times the half-life λ for the exponential weights. We can measure relative precision as the ratio of the variances of the estimates, θ^2. The exponentially weighted estimated has variance θ^2 times the variance of the equally weighted estimate. A useful approximation of this variance ratio is $\theta^2 \approx 0.35\ T/\lambda$, provided $T/\lambda > 4$.

The analysis and approximate variance ratio can serve as a guide when we are trying to balance a rough sense of the benefits from exponential weighting against an approximate loss in efficiency from exponential weighting.

4 References

Bollerslev, Tim, 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* 31, 307–327.

Chow, Gregory C., 1983, *Econometrics* (McGraw-Hill, New York, NY).

Fama, Eugene F., 1965, The behavior of stock market prices, *Journal of Business* 38, 34–105.

Hayashi, Fumio, 2000, *Econometrics* (Princeton University Press, Princeton, NJ).

Hentschel, Ludger, 1995, All in the family: Nesting symmetric and asymmetric GARCH models, *Journal of Financial Economics* 39, 71–104.

Lo, Andrew W., and A. Craig MacKinlay, 1988, Stock market prices do not follow random walks: Evidence from a simple specification test, *Review of Financial Studies* 1, 41–66.

Muth, John F., 1960, Optimal properties of exponentially weighted forecasts, *Journal of the American Statistical Association* 55, 299–306.

A Weights in Continuous Time

The main text discusses the effects on precision if we apply exponential weights in discrete time. For completeness and comparison, this appendix provides related analysis for continuous time.

A.1 Continuous Time Weights

In continuous time, exponentially declining weights are

$$w(t) = \alpha(1 - e^{-\alpha T})^{-1} e^{-\alpha t}. \tag{11}$$

These weights are scaled so they integrate to 1 over the interval 0 to T. As α goes to zero, the weights converge to equal values $1/T$.

The variance of the weighted mean is

$$\begin{aligned} \int_0^T w^2(t)\,\sigma^2\,dt &= \alpha^2(1 - e^{-\alpha T})^{-2} \int_0^T e^{-2\alpha t}\,\sigma^2\,dt \\ &= \frac{\alpha}{2}(1 - e^{-\alpha T})^{-2}(1 - e^{-2\alpha T})\,\sigma^2 \\ &= \frac{\alpha}{2}\,\frac{e^{\alpha T} + 1}{e^{\alpha T} - 1}\,\sigma^2. \end{aligned} \tag{12}$$

The last step follows from $(1 - e^{-2\alpha T}) = (1 - e^{-\alpha T})(1 + e^{-\alpha T})$.

The variance relative to the variance of the equally weighted estimator, σ^2/T, is

$$\theta^2(\alpha, T) = \frac{1}{2}\,\alpha T\,\frac{e^{\alpha T} + 1}{e^{\alpha T} - 1} \tag{13}$$

$$= \frac{T}{\tau(\alpha, T)}, \tag{14}$$

with

$$\tau(\alpha, T) = \frac{2}{\alpha}\,\frac{e^{\alpha T} - 1}{e^{\alpha T} + 1}. \tag{15}$$

As for discrete time, we can show that

$$\lim_{\alpha \to 0} \tau(\alpha, T) = T, \tag{16}$$

which means that the variance ratio approaches 1 as the weights converge to equal weights.

The half-life of the exponential weights is

$$\lambda = \frac{\ln(2)}{\alpha}, \tag{17}$$

so that

$$\alpha = \frac{\ln(2)}{\lambda}. \tag{18}$$

Substituting equation (18) into equation (13) yields

$$\begin{aligned}\theta^2(\lambda, T) &= \frac{\ln(2)}{2}\,\frac{e^{\ln(2)T/\lambda}+1}{e^{\ln(2)T/\lambda}-1}\\ &= \frac{(2^{T/\lambda}+1)}{(2^{T/\lambda}-1)}\,\frac{\ln(2)}{2}\,\frac{T}{\lambda}.\end{aligned} \tag{19}$$

Although the relative precision is not linear in αT or T/λ, it only depends on the product αT or the ratio T/λ, not the separate values.[7]

From equation (19), we can approximate the relative precision as

$$\begin{aligned}\theta^2(\lambda, T) &\approx \frac{\ln(2)}{2}\,\frac{T}{\lambda}\\ &\approx 0.35\,\frac{T}{\lambda}\end{aligned} \tag{20}$$

for large values of T/λ. This is the same approximation we derived in discrete time. In continuous time, the approximation is more obvious. Also, in continuous time the approximation improves for very large values of T/λ. As before, this approximation does no work well for small values of T/λ and does not obey the lower bound $\theta^2 \geq 1$.

Alternatively, we can linearize θ^2 around $T/\lambda = 0$ and find

$$\begin{aligned}\theta^2(\lambda, T) &\approx 1 + \frac{\ln(2)^2}{12}\left(\frac{T}{\lambda}\right)^2\\ &\approx 1 + 0.04\left(\frac{T}{\lambda}\right)^2.\end{aligned} \tag{21}$$

By construction, this approximation works well for small values of T/λ. It also obeys the lower bound $\theta^2 \geq 1$. Unfortunately, this approximation becomes unreliable even at moderate values of T/λ, say $T/\lambda > 5$. Higher-order terms in the expansion do not materially extend the accurate range and seem impractical for mental calculations. Generally, we should be more concerned about the loss of precision for larger values of T/λ. In those cases, the approximations in equations (10) and (20) are more useful.

---

[7] This is not exactly true in discrete time because the discrete-time variance ratio converges to T, not infinity, as T/λ goes to infinity.